ChinaXiv合作期刊

学术图表知识发现技术框架及研究进展*

■ 丁培
深圳大学图书馆 深圳 518060

摘 要：[目的/意义] 科技资源深度融合背景下，学术图表知识发现是提供除文本知识发现外新的知识发现方式，是完善文献知识发现的重要一环，可提升科研人员科学发现及知识创造效能，推动数字图书馆知识服务升级。[方法/过程] 梳理学术图表知识发现的演进脉络，详细论证其技术框架内容，证明学术图表知识发现技术逐步成熟。结合学术图表知识发现应用服务，论证学术图表知识发现在科技创新多方面有广阔应用空间。[结果/结论] 展望学术图表知识发现未来，我们需要：重视学术图表知识发现，将其融入文献知识发现体系内；完善学术图表语义知识组织体系，构建专门的学术图表语义知识库；开发新型学术图表知识发现应用。

关键词：学术图表 知识发现 知识组织 信息抽取
分类号：G254
DOI：10.13266/j.issn.0252-3116.2021.23.015

chinaXiv:202304.00405v1

1 引言

科技信息资源深度融合背景下，全新的数据密集型科学发现成为科技创新生态。人工智能及深度学习技术取得突破性进展，这给支撑新生态的知识发现服务带来新变革和新要求。数字图书馆领域中，以文献知识发现为核心的知识发现服务转向对象多源异构、内容组织细粒度、跨类型语义关联、机器可理解及机器发现知识的趋势愈发明显，以学术文本为中心的传统知识发现面临异构载体及新服务挑战。学术图表是科技文献中用于内容描述、论点支撑、数据对比的各类图表数字对象。N. Siegel 采集分析 arXiv 和 PubMed 中550万篇科技文献发现 arXiv 的 PDF 论文仅20%没有学术图表，而 PubMed 的 XML 文件仅10%没有学术图表[1]。生物医学领域，几乎每篇期刊文献都包含学术图表，它们比任何类型信息更能代表医学文献中的证据内容[2]。相关研究发现学术图表提供比文本更多的信息，利用学术图表能有效提高用户发现文献的效率[3]。P. Lee 发现影响力越大的论文往往包含更多学术图表[4]。学术图表支撑科研再利用，解释文献重要研究内容，是科技文献资源与科技数据资源融合交叉点，是科研人员重视的科技知识载体。

长期以来，由于学术图表视觉与文本特征共存、表现形式多样、信息抽取复杂等因素，机器理解学术图表停留在弱语义层次，致使学术图表难以有效融入现有文献知识发现体系内。未来学术知识服务体系需要细粒度知识组织、基于语义的知识关联、面向全类型资源的知识发现以及能有效支持智能问答、意图精准刻画的认知计算。作为典型异构学术对象，研究学术图表知识发现对完善文献知识发现体系、推动科技资源深度融合、促进非文本型数据知识发现、创新数字图书馆知识服务有积极意义，也十分必要且迫切。

本文以“图像 表格”“信息抽取”“科技文献 论文”为核心检索词，并扩展“图像识别 表格识别”“图像标注 表格标注”“知识发现”“命名实体识别”“图表关系抽取”等相关概念，分别在 Web of Science、Scopus 及 CNKI 数据库中进行主题检索，数据检索时间截止到2021年8月。基于文摘阅读筛除不相关论文，确定密切相关文献85篇。在此基础上，基于参考文献扩展相关文献135篇，共同形成本文的研究基础。本文梳理学术图表知识发现演进脉络，并以技术框架及流程为骨架综述各技术点的研究分支及进展，最后展望学术图表知识发现下一步研究。

2 学术图表知识发现演进脉络

学术图表发现经历对象发现到知识发现的演变。对象发现是指从科技文献中抽取、组织、检索发现学术

* 本文系广东省哲学社会科学规划学科共建项目“支持深度知识发现的文内数据与文献关联研究”（项目编号：GD18XTS07）研究成果之一。
作者简介：丁培，馆员，博士研究生，E-mail：peid@szu.edu.cn。
收稿日期：2021-06-08 修回日期：2021-09-12 本文起止页码：136-148 本文责任编辑：杜杏叶

图表的过程。学术图表对象发现又经历学术图表对象的简单发现——学术图表对象关联文献发现——学术图表对象的多维发现三个阶段:①学术图表对象的简单发现阶段,学者们关注如何从科技文献内提取出单一的学术图或学术表,并采用元数据方式组织学术图表的简单信息,提供基于关键词的学术图表发现;②学术图表对象关联文献发现阶段,在前期研究基础上,研究者们将学术图表上下文内容也作为学术图表发现的重要信息来源,建立学术图表和所在文献的关联,尝试将学术图表融入科技文献发现系统中。与此同时,这一时期学术图像分类研究大量涌现,学术图像分类组织成为此阶段新的特色;③学术图表对象的多维发现阶段,部分大型数字资源商(如 Pubmed、CNKI)参与到学术图表对象发现,他们探索更多的发现方式,如利用学术图像的图像特征实现图-图发现,尝试利用自然语言处理技术、机器学习算法等自动化抽取学术图表、学术图表文本内容及学术图表所在文献的元数据来解决海量学术图表信息发现,尝试引入语义知识组织体系(如主题词表)来实现语义扩展发现。表 1 总结了学术图表对象发现不同阶段的相关研究与实践:

表 1 学术图表对象发现不同阶段的相关研究与实践

不同阶段	发现内容	主要应用技术	组织方式	发现方式	相关研究及实践	实践时间/年
学术图表对象的简单发现	学术图表标题、注释、学术表条目、学术图像图例	学术图表对象获取及文本获取	元数据组织	关键词发现	TINTIN[5] FigSearch[6]	1997 2004
学术图表对象关联文献发现	学术图表标题、注释、文献标题、学术图表上下文、学术图像类型	学术图表对象获取及文本获取/人工标注	元数据组织/图表分类组织	关键词发现/图表类型发现	CSA Ilustrata[7] TableSeer[8] Yale Image Finder[9]	2006 2007 2008
学术图表对象的多维发现	学术图表标题、注释、学术图表上下文、学术图像类型、学术图表主题、文献标题、作者、相似学术图表	学术图表对象获取及文本获取/人工标注/图像自动分类/文本自动标注	元数据组织/图表分类组织/主题词表	关键词发现/图表类型发现/主题词发现	Biomedical Figure Search[10] Pubmed Central[11] CNKI[12] Open-i[13] FigureSeer[14]	2010 2011 2011 2014 2016

chinaXiv:202304.00405v1

对象发现一定程度上满足科研人员查找非文本型学术资源的需求,但其仅揭示学术图表的显性信息,并未识别和揭示学术图表内隐藏的其他知识。此外在对象发现中,学术图表与文本发现割离,不利于两者知识互通、融合。近年来,机器视觉识别技术、文本深度挖掘技术、语义组织技术快速发展与成熟,学术图表发现从仅发现学术图表对象逐渐走向发现学术图表隐藏知识的学术图表知识。

知识发现(Knowledge Discovery in Database, KDD)是基于数据库的知识发现,它是从数据中识别出有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程[15]。学术图表知识发现是从海量文献内海量学术图表数据中自动构建、发现新的知识模式的过程。这一过程并非是人工演绎、归纳和推理过程,而是机器学习过程。学术图表存在着文本信息表示和视觉信息表示的双模态特征,它的双模态意味着学术图表知识发现需要统计学的机器学习算法、强大的数据库技术支持、融合语言学词汇及句法特征处理文本和训练知识模式、以及基于机器视觉识别挖掘学术图表视觉特征中隐藏的知识模式。

相比于学术图表对象发现,学术图表知识发现在三方面突破:首先,学术图表知识发现不再割裂学术图表和科技文献文本。通过挖掘学术图表中的显性及隐性知识,并基于数字知识模式表示消除学术图表和学术文本间的模态隔阂,学术图表知识发现实现知识层面上的跨模态发现。计算机真正将学术图表理解为科技文献的知识组成部分;其次,知识发现面向海量数据处理,因此自然语言处理、图像自动分类、文本自动分类、自动语义标注、信息抽取等技术是学术图表知识发现的重要支撑。语义知识组织是学术图表知识发现的主要组织方式,协助多源异构系统检索和细粒度内容发现;第三,知识模式发现是学术图表知识发现的重心。学术图表知识发现将在本体等领域知识组织体系和人工标注语料的基础上,融合视觉对象识别、术语抽取、语义标注、关系抽取等技术,对复杂知识实施自动抽取及建模。

3 学术图表知识发现技术框架

知识发现具有流程化特点。文本知识发现技术框架包括自由文本预处理、文本表示和编码、文本分类或聚类、信息抽取/知识抽取 4 个部分。学术图表知识发现同样由数个关键技术节点构成组成技术框架。基于知识发现基本流程,结合学术图表自身特性,确定学术图表知识发现的 4 个关键技术节点:学术图表对象及文本的识别与获取、学术图表信息表示与建模、学术图

ChinaXiv合作期刊

表分类和文本分类、学术图表信息抽取。图 1 展示了各技术点的流程关系：

图 1 学术图表知识发现技术框架

3.1 学术图表对象及文本的识别与获取

3.1.1 学术图表对象识别与获取

学术图表知识发现首先要识别、定位、获取科技文献中的学术图表，并建立学术图表和周围文本间的联系。规范化标记格式（如 HTML/XML 格式）和 PDF 格式是目前主流的两类科技文献格式，学术图表识别任务在两类格式上所需技术存在差异。图 2 展示了学术图表对象识别与获取在不同格式中的技术区别：

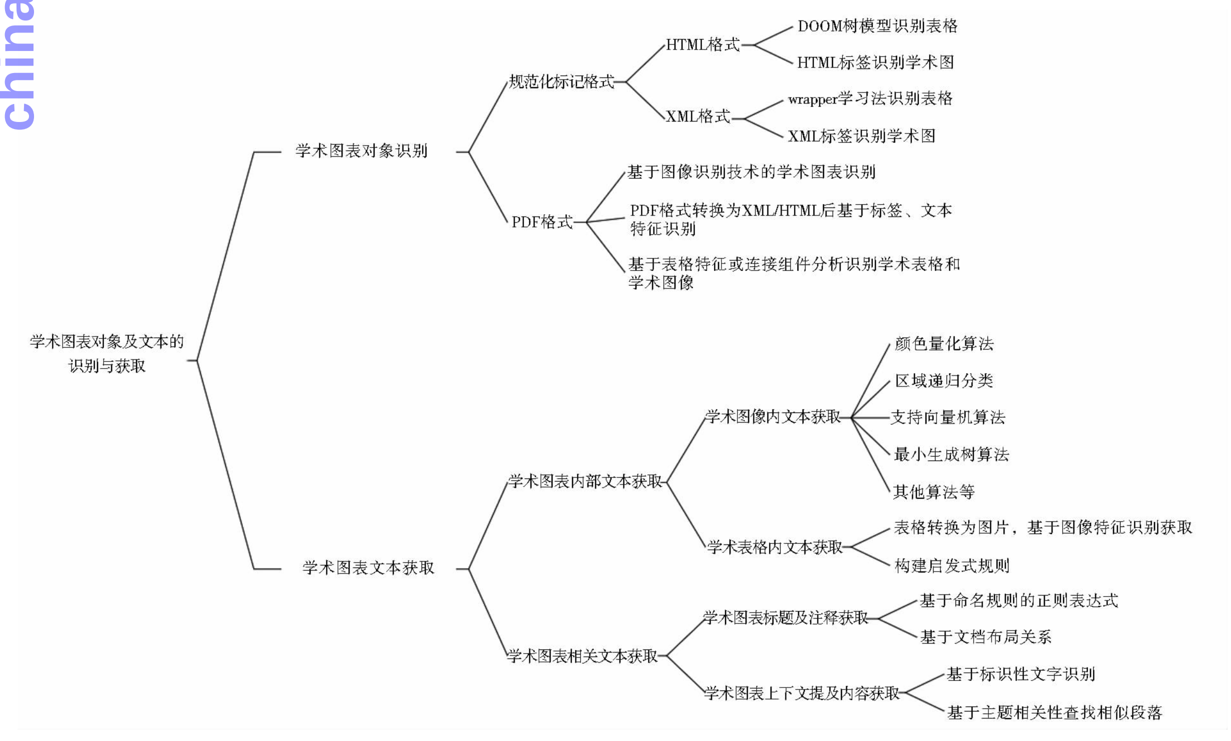


图 2 学术图表对象及文本的识别与获取技术路线

chinaXiv:202304.00405v1

chinaXiv:202304.00405v1

HTML 兴起之初,研究者基于 ASCII 文件、光学字符识别或特殊的制表符标记,构建表格 DOOM 树模型来识别 HTML 文档内的学术表格[16]。XML 时代,学术图表数据与 XML 文档分开存储,因此学术图表能直接获取,只需要根据 XML 中学术图表标签及路径来建立学术图表和 XML 文本文档间的关联。针对学术表格直接存在于 XML 文档中的情况,需要分析表格结构,结合特定标签,使用 wrapper 学习方法获取表格内容,并重新组合表格[17]。

PDF 文档内图像识别研究较早[18]。PDF 文档中图像提取相对更难。图形数据通常会以 raster 栅格(PNG,JPEG)或 vector formats 矢量格式(SVG,EPS)嵌入于 PDF 文档中。研究人员采用两类方法来识别及分离 PDF 中图片:①基于图像特征的图－图识别。首先通过扫描等方式将 PDF 整个转为图片,随后基于位图分割技术[19]、区域分类[20]、基于连接组件[21]的方法识别 PDF 中的学术图像。②格式化标签识别,即将 PDF 转换为结构化的 XML/HTML 格式,然后基于标签识别提取图像。例如 Apache PDFBox[22]、PDFMiner[23]、Xpdf[24]和 Poppler[25]等工具将 PDF 文档转换为结构化的 XML/HTML 格式,并提取文档中的图形。这些工具提取矢量格式的图像时,只能识别图形中的单个组件,例如直方图的一个条形段,而不是提取整个图像。针对这个问题,部分研究者提出基于正则表达式(启发式)来识别图像标题,基于标题位置,利用聚类算法来识别特定图像[26],或利用分类算法排除无关的矢量图像[27－28],从而达到提取整个图像的目的;P. Y. Li 等将文本内容与 PDF 文件的图形内容分开,利用连接组件分析检测图像,并基于 PDF 的布局信息恢复图像标题并建立与图像间的关系[29]。

PDF 文档中学术表格识别获取分三种技术路线:①使用第三方软件将 PDF 转换为 XML 或 TXT 格式,基于标签及文本特征抽取表格[30]。②针对以图片方式存储于 PDF 中的表格,引入图像识别技术,基于图像特征,经过灰度变换、图像平滑、边缘检测、二值化和倾斜矫正等步骤分离并获取表格[31]。③基于 PDF 表格特征(如文字栅格、框线等),通过解析算法,直接在 PDF 中获取表格文本,实现表格形态的还原[32]。相关研究开发 Tabula[33]、TEXUS[34]、TAO[35]等表格提取工具。

3.1.2 学术图表文本识别与获取

(1)学术图表内部文本获取。学术图表内部文本指学术图像中的图例、图注、图像内文字等内容。J. Sas[36]、F. Böschen[37]总结学术图像中文本提取的通用步骤包括图二值化处理、图像特征矢量计算、应用连接组件标记、OCR 识别、特殊字符过滤等。

为解决通用方法准确率不稳定的问题,研究者们使用不同方法从特定学术图像提取图内文本。如在制图地图中应用颜色量化算法,使用形态学算子和 OCR 来检测并分离文本[38];使用垂直和水平投影直方图分析,将直方图的各区域递归分类为文本和非文本[39];使用基于几何、区域、示例和轮廓等相关特征,采用支持向量机分类算法,从生物医学出版物图像中自动检测识别文本[40];利用深度学习模型和 OCR 识别从生物学领域的路径图中获取分子实体及其相互作用的文本内容[41]。

表格内文本抽取研究相对成熟,有两类方法:①将表格转为图片,基于布局、线条、文本位置、单词间距、文字大小等特征,按照图片内文本抽取的步骤,采用贝叶斯分类算法或者树形遍历算法,从图片内抽取文本内容[42];②基于规则,构建启发式或模板,识别横纵轴标签及数值,抽取表格实体并重构关系[43]。

(2)学术图表相关文本的获取。Y. Hong 研究发现,若不参考上下文提及文本,研究人员理解学术图表将丢失 30% 的信息内容,因此理解和发现学术图表应结合学术图表和上下文提及文本[44]。获取学术图表上下文信息需要保证尽可能找到学术图表涉及的文本内容,也应尽量少引入无关的文本信息。其中学术图表标题、注释及正文中学术图表上下文提及内获取是主要研究点。

学术图表标题及注释获取可分为基于规则和基于布局关系两种方式:①基于规则的方法利用特定字段或基于命名规则的正则表达式来获取学术图表标题及注释内容。如利用 < caption >、< table － note >等字段获取 XML 中的标题和注释内容。PDF 文档中标题及注释抽取可基于命名规则,利用正则表达式来抽取[45]。基于规则的方法需要过滤器来筛选噪音结果。如仅选择以分号、句号、冒号为结尾的短语;或选择粗体或斜体;或选择字体与后面不一致的短句;或聚类不同描述符组,选择最多数量的组为唯一标识[46]。②基于布局关系的方法利用学术图表和学术图表标题、注释在文档布局上的对应关系,使用图像识别技术抽取图下或者表上的学术图表标题[29]。例如 C. Clark 和 S. Divvala 将每页 PDF 分解为标题、正文、图形文本和图形等不同区域,构建图形重叠、垂直文本、宽间隔文本、行宽等启发式对标题、学术图表注释、正文文本分类[28]。

学术图表标题和学术图表本身匹配也是重要研究

ChinaXiv合作期刊

问题。XML 格式论文文档通常会提供学术图表的引用 ID,基于 ID 名称可建立学术图表标题和学术图表本身间的对应关系。PDF 文档内多数需要基于不同的学术图表和标题布局,综合考虑标题和学术图表的 1-to-1、N-to-N、N-to-M 关系,利用相关算法来确定对应关系[29]。

学术图表上下文提及内容获取有两种方法:①方法基于标识性文字来识别明确引用学术图表的句子或者段落,如 fig、table 等关键词[47]。②方法以学术图表标题或明确引用学术图表的语句或段落为基准,基于主题相关性来查找与之最相似的句子或段落[48-49]。

综合而言,学术图表对象及文本的识别与获取任务在不同文献类型中发展出不同的技术路线。学术图像和学术表格的识别在现有技术支持下能获得不错的效果。学术图表文本识别中的上下文提及内容获取是一个难点,需要在覆盖率和准确率上取得平衡。

3.2 学术图表信息表示及建模

学术图表知识表示是指将描述学术图表的自然语言文本以及学术图表所展示的图像视觉信息变为计算机可处理的数字知识表示模式。学术图表涉及三类信息表示,分别是学术图表文本表示、图像视觉特征表示、图像标注文本表示。如图 3 所示:

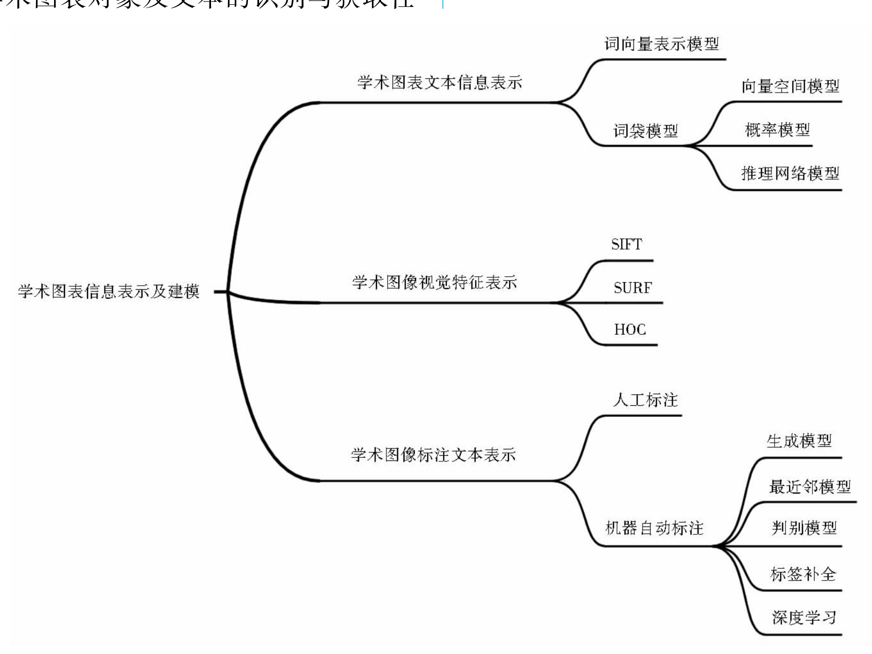


图 3 学术图表信息表示及建模技术概览

3.2.1 学术图表文本表示

在文本知识发现中,基于离散的词表示为基础的文档表示模型是的文本表示模型,其中词袋模型(Bag of Words)是最常见的文本表示方式,在词袋模型上进一步衍生出向量空间模型、概率模型[50]和推理网络模型[51]等表示模型。TF-IDF 是传统空间向量模型中用于特征权重计算的常见方法,分布式词嵌入表示(word embedding)是神经网络模型支持下产生的热门词向量表示模型[52]。在学术图表的标题、注释及上下文可延续使用上述文本表示方法。

3.2.2 学术图像视觉特征表示

图像视觉特征表示是利用不同形式的特征表示描述图像的视觉内容的过程,此过程是让机器理解图像的基本单元。基于视觉特征表示的图像检索又称为基于内容的图像检索(Content Based Image Retrieval, CBIR)。视觉特征表示过程大致分为三个步骤:区域选择、特征表示、特征聚类。

区域选择早期采用固定划分的方式,此方式简单但破坏了图像的视觉内容。图像分割是研究最多的区域选择方法,其最终目的是将分割后的像素归属于一个对象,包括有监督[53]、弱监督[54]及无监督[55]的分割算法。事实上,图像分割不仅是底层图像处理问题,同时是对象理解问题。目前自动图像分割在特定领域表现不错,但在通用领域上欠佳。显著点选择是对象级分割难以提升准确率的优化区域选择方式,其原理是选择图像中具有显著特征的点来表示图像区域[56]。

区域选择后需要从确定的图像区域内提取出图像视觉内容的特征信息,如常见的颜色、纹理、形状和空间关系等,并在特征提取后通过特定描述符来表示图像视觉的局部对象,这就是图像特征表示,也称为视觉

chinaXiv:202304.00405v1

单词袋(Bag of Visual Words,BVW)。SIFT(Scale Invariant Feature Transform)[57]、SURF(Speeded-up Robust Features)、HOG(Histogram of Oriented Gradients)[58]等是应用较多局部特征表示方法。尽管提取的图像视觉特征信息能直接用于图像检索,但存在向量维度过高的问题,需要降维处理。降维方式有主成分分析[59]、奇异值分解[60]、局部敏感哈希[61]等。

在深度学习技术支持下,近年来有诸多研究尝试使用视觉语义嵌入学习[62]、共识感知视觉语义嵌入[63]、图注意力[64]等方法挖掘图像和文本间的潜在语义结构信息,计算图像视觉特征表示和文本表示的相似性,从而实现基于图像的文本检索或基于文本的图像检索。它们致力于将图像视觉表示和文本表示统一在一个空间上,但当下此类技术未能平衡全局特征和局部区域特征的关系,暂时未应用到更多的跨模态任务,如图像字幕和视觉问答中。

3.2.3 学术图像标注文本表示

单纯的图像视觉特征表示无法让机器理解图像高级语义概念,这导致机器与人理解图像上的语义鸿沟。图像标注正是为建立机器理解的视觉特征与人理解的文本内容间映射而产生研究主题。学术图像标注采用人工或机器自动学习的方式,将学术图像的低层视觉特征表示为高级语义的标注文本内容,这些与学术图像关联的标注文本可作为计算机理解学术图像的数字知识表示[65]。主流的 5 种图像自动标注方法包括基于生成模型、基于最近邻模型、基于判别模型、基于标签补全、基于深度学习[66]。其中基于深度学习算法的图像自动标注是近年的研究热点,涉及到的模型包括深度神经网络、卷积神经网络、循环神经网络、长短期记忆网络及堆栈自动编码等[67]。这些自动标注方法大多实验于一般图像或网络图像,而学术图像领域,目前主流标注方式依旧是人工标注,发展了 Quick Annotator[68]、DicomAnnotator[69]等半自动或众包标注工具。

学术图表的双模态导致学术图表信息表示上的割裂。学术图像标注文本表示尝试修复文本表示和视觉特征表示间的割裂,但由于学术图像标注所需初始标注知识库缺乏,同时受制于学术图表为核心对象的知识单元语义表示模型尚未完善,导致学术图表自动语义标注技术未能实现大规模应用。图像视觉表示和文本表示统一到同一空间计算是值得关注的技术,需关注其在全局空间和局部对象的结合以及视觉语义推理上的进展。

3.3 学术图表分类和学术图表文本分类

学术图表分类及文本分类是学术图表检索等学术图表知识发现应用的基础。文本分类是使用预先的知识分类框架或者规则,基于逻辑模型(例如决策树)、概率模型(例如朴素贝叶斯)、几何模型(例如支持向量机)等对文本进行分类处理[70]。

学术图表文本分类分为两个子任务:一是学术图表上下文分类,例如将上下文分为简介、方法、结果和讨论等,其主要用途是文本摘要形成;二是学术图表内文本分类。学术图像中部分文本有明确含义,如图例、x 轴标签、y - 轴标题等,可以对它们实施分类。J. Poco 等构建一个专门的学术图像文本分析管道,通过文字检测、OCR 识别、词合并、文本分类等步骤实现学术图像内文字编码的逆向解析,并将其分类为不同实体类型[71]。学术表格文本分类则关注其在文献内使用功能。如 S. Kim 将科学论文内的表格分为背景、系统/方法、实验三类以及评论、比较两个功能类[72]。

学术图像分类已有大量研究。学术图方面,相关研究融合图像低层特征及文本特征,基于支持向量机[73]、卷积神经网络算法[74]、多样性密度算法[75]等机器学习模型,实现部分学术图像的自动分类,如条形图、饼图、折线图、射线图等。复合图作为常见的一种学术图像类型,其识别及子图类型分类是当下研究热点之一。

复合图识别分为基于文本特征、基于视觉特征、基于混合特征三种方法。文本特征指复合图中及图注内的文本标签内容,例如复合图的拼接处以及图注释中"A.""b:""(c)"等标识,其标识格式一般为序列符号+分隔符号。研究者利用这些文本特征,使用正则表达式[76]或支持向量机算法[77]识别学术复合图。基于视觉特征的复合图识别依靠的是图像的布局信息,例如子图间的空白。研究者基于复合图视觉特征借助分界线探测[78-79]、子图连通域探测[80]、图像强度统计[81]等技术识别复合图。

子图类型识别是多标签分类任务,分两种方法:①分割复合图为子图,随后基于单一图的分类算法识别子图类别标签[82];②创建多标签学习模型,基于复合图说明文本及复合图视觉特征,直接从复合图中识别子图类别[83]。

学术表格分类任务研究较少,主要从表格形态、用途等维度对表格分类,如 Tabex 工具识别 Web 表格,并将其分为垂直列表、水平列表、日历、窗体等[84]。

学术图表文本分类和学术图表分类均能有效提高学术图表信息抽取的效果。当下学术图像内的文本分类任务依旧局限在文本功能层面,未来可以结合图像

chinaXiv:202304.00405v1

ChinaXiv合作期刊

类型,对文本进行语义深度分类,探究图像类型与图像文本间的语义关联,例如流程图中的文本表示流程步骤,树形图的文本存在的上下级关系。由于不同领域中学术图像表现类型不一以及复合学术图的大量存在,学术图像类型识别尚无法做到全部图像类型覆盖。

3.4 学术图表信息抽取

信息抽取是知识发现最重要的一环,它从非结构化数据中抽取出结构化信息以获得知识初始模式。命名实体识别及关系抽取是文本信息抽取的两个核心过程。

学术图表信息抽取包括两大方面:①科技文献中与学术图表相关的文本信息抽取,此部分的技术路线即传统的科技文献文本信息抽取的技术,已有诸多论文论述,本研究不再详述;②学术图表本身的信息抽取,其又分为学术图像内实体识别及标注、学术图表关系抽取两个分支。图 4 展示了学术图表信息抽取的主要技术点及其主流方法:

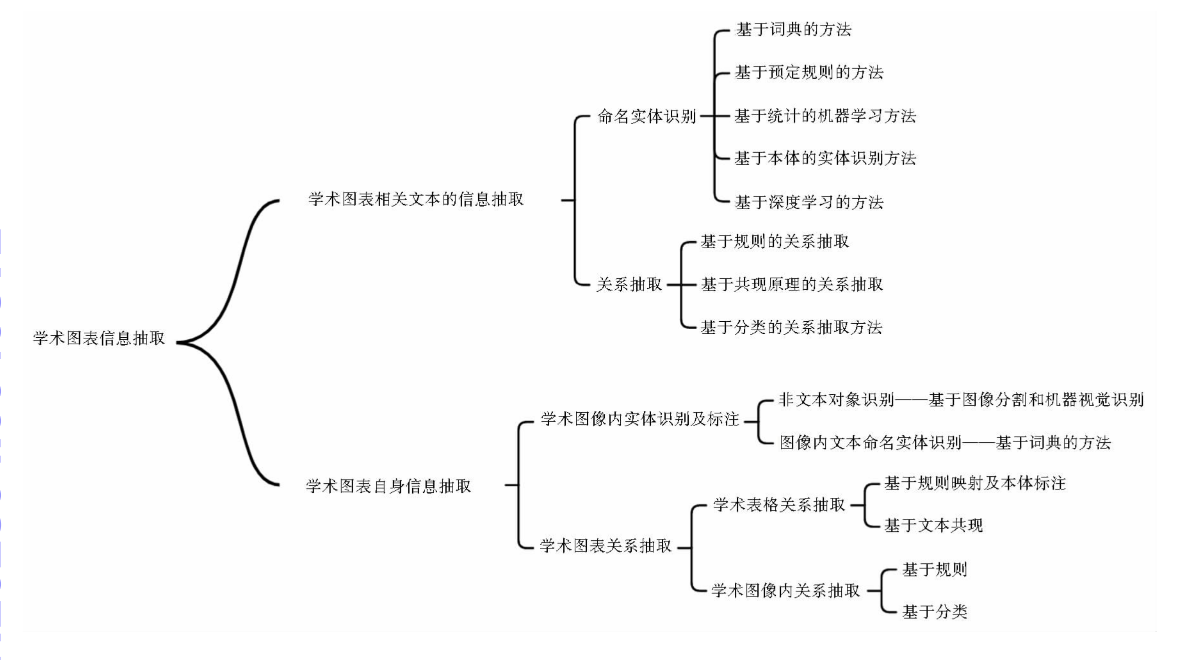


图 4 学术图表信息抽取技术路线

3.4.1 学术图像内实体识别及标注

学术图像内实体识别及标注涉及学术图像内非文本对象识别与标注、学术图像内文本命名实体识别。

学术图像内非文本对象识别与标注是基于图像分割和机器视觉对象识别,从照片、医学图像、成像图等类型的学术图像中发现科研对象,并建立对象的边界与类别。研究者们在医学、生物、农业等领域开展特定类型图像的非文本对象识别与标注。结构化文本图片发现系统(Structured Literature Image Finder system,SLIF)关注生物文献中的显微镜成像图,通过机器视觉识别的方法来发现成像图中的基因、蛋白质对象,并标注概念[85]。Human Brain Project 项目识别大脑成像图片的特定区域对象,并将其与受控词表中的概念关联[86]。EMAP (the Edinburgh Mouse Atlas Project)利用解剖学词表概念对老鼠胚胎的 3D 图片和 2D 组织切面进行标注[87]。农业领域的研究者们基于卷积神经网络等深度学习算法识别并分类学术图像中的不同植物的不同病虫害,在小范围数据集中取得不错效果[88-89]。

学术图像内文本命名实体识别通过识别学术图像中的文本对象,基于图像表达内容来实施命名实体识别。如 T. Kuhn 等识别医学文献中凝胶图片中标签,对基因、蛋白质等对象进行命名实体识别,正确识别基因/蛋白质实体达到 65.3% 左右[90]。

3.4.2 学术图表关系抽取

(1)学术表格关系抽取。研究者基于表格形式特征,抽取学术表格文本并借助本体或语义映射关系来抽取学术表格内文本关系。Z. Q. Zhang 提出了一种增量的、互递归的、弱监督学习的一维表数据自动语义标注方法 TableMiner,利用上下文信息和部分列数据初步得出列头对应的类和单元格数据在 FreeBase 知识库中对应的实体,并抽取实体关系[91]。H. P. Cao 等借助本体工具,利用规范化的观测术语、实体对象,基于表格对应关系,将观测数据表格转化为可理解的事件[92]。C. S. Bhagavatula 等构建了实体链接系统 TabEL,该系统通过考察单元格短语与候选实体在维基百科文档和表格中的共现情况来确定列类型和列关系[93]。

chinaXiv:202304.00405v1

(2)学术图像内关系抽取。学术图关系抽取建立在图内文本、对象、数值提取的基础上,可基于规则或分类的方式抽取知识关系。A. Kembhavi 等引入一种图解析图(Diagram Parse Graphs,DPG)方法,识别文献中视觉插图(如食物链图、大气循环图等)中的图元素,并建立元素间的语义关系[94]。P. Lee 等提出从系统树图中提取信息的新方法,可以实现科学文献中系统树图自动识别,并基于层级规则,提取树结构的关键成分,重建树,恢复树的层次关系[95]。何英研究科技文献中的柱形图的检测、分割、信息提取,基于 CNN 卷积神经分类器,从生物文献中的柱形图中抽取大豆基因和表型相关的数据,挖掘并建立基因-表型-育种时间-表现水平数值间的关系[96]。

学术图表信息抽取是综合性任务,一方面它需要学术图表信息表示及学术图表分类为其提供基础信息,另一方面抽取任务要深度融合语义信息。现有信息抽取研究实践通过借助领域词典或自定义语义关系可实现特定学术图表类型中的部分语义信息抽取。若能够建立完善的学术图表语义知识组织体系,并将其与领域知识组织体系结合,必将获得更精准的学术图表信息抽取。

chinaXiv:202204.00405v1

4 学术图表知识发现应用服务

知识服务应用是学术图表知识发现的落脚点。目前,学术图表知识发现主要应用于三大方面,分别是学术图表检索、学术图表自动摘要、图像视觉问答。

4.1 学术图表检索发现

学术图表检索是最广泛的学术图表知识发现应用。它涉及学术图表识别、学术图表分类及学术图表标注等知识发现技术。例如 CSA llustrata 学术图表检索识别抽取文献中的表格、图片等数据,通过"深度索引"方法人工标引元数据建立独立索引数据库,继而提供基于关键词的学术图表检索服务。

随着知识发现技术持续深入,学术图表检索呈现新的特点。表现在:①学术图表分类中更多采用机器学习的自动分类方法;②利用语义标注技术提供基于本体推荐的语义相关术语来优化查询;③使用文本自动分类及相似度计算形成学术图表的自动摘要内容。NIH 开发的科研图片数据库 Open-i 平台是代表之一。该平台综合来自 PMC、Medpix、USC Orthopedic Surgical Anatomy、Images from the History of Medicine (NLM)、Indiana U. Chest X-rays 等的科研图片,其中 PMC 的科研图片是科技文献内的学术图像。Open-i 提供关键词、Mesh 主题词检索以及以图找图的发现方式,并采用图像自动分类、图像语义标注及图像文本自动分类等相关知识发现技术。

4.2 学术图表自动摘要

学术图表文本摘要能够帮助科研人员快速了解学术图表含义而不用阅读论文全文,同时学术图表摘要配合学术图表检索能单独提供知识发现服务。文本摘要应用的主要知识发现技术包括学术图表上下文提及内容获取、文本分类、信息抽取等。文本摘要分抽取型摘要和抽象型摘要两类:抽取型摘要基于语句语义关系定义及预训练直接从原目标文档中抽取已有片段来构建摘要。抽象型摘要则灵活抽取事实对象或语句,生成的摘要可能含有原文中并不存在的词或句子。

目前学术图表摘要以抽取型摘要居多。根据摘要形成使用的方法类型,分为有监督学习和无监督学习。其中有监督学习需要先训练样本,如 S. Bhatia 分别使用朴素贝叶斯和支持向量机的分类算法,根据文章句子与学术图表标题之间的相似度,抽取相关句子形成学术图表摘要内容[49]。S. Agarwal 等开发图形摘要系统 FigSum,从医学文献中抽取出图形的结构性文本摘要,并将文本分类为简介、方法、结果和讨论[97]。无监督学习不需要预先训练,而是机器自动学习分类。N. Saini 等采用多目标优化(Multiobjective Optimization, MOO)方法构建了无监督的学术图自动摘要系统 MOOFigSum[98] 和 FigSum++[99],能自动为论文内每一个学术图表生成摘要。J. Chen 等采用无监督的分层多模态 RNN 模型生成文本+图像的多模态新闻摘要[100]。

4.3 图像视觉问答

图像视觉问答(Visual Question Answering)融合计算机视觉及自然语言处理两大人工智能领域技术,是当下的研究热点。其形式是通过向机器输入图像以及关于图像内容的自然语言形式问题,机器反馈自然语言形式的回答。这其中涉及图像对象识别、图像标注等知识发现技术。

目前视觉问答主要集中于自然图像理解领域,研究者们提出基于图像特征融合、基于实体注意力、基于多步推理、基于引入知识、基于关系建模等多种视觉问答方法[101]。学术图像领域,研究者开展特定类型图像的视觉问答研究及学术图表视觉问答数据集构建等研究。A. Kembhavi 等通过引入图解析图注意力模型方法,抽取文献中视觉插图元素及插图文本,建立元素与文本间的对应语义关系,基于长短记忆神经网络学习算法解析语法,构建视觉插图知识问答系统[95]。K. Kafl 在视觉问答基础上提出一个专门用于文献中条形

ChinaXiv合作期刊

图的数据检索和数据推理的视觉问答数据集 DVQA[102]。微软研究构建了一个可用于学术图表问答的数据集 FigureQA[103]，含 18 万张垂直条形图、水平条形图、折线图、虚线图以及饼图，有超过 200 个问题及答案，为开发功能更强大的学术图表视觉问题回答和推理模型提供参考。类似数据集还有 LEAF-QA[104]。整体而言，学术图像的视觉问答应用前景广泛，但还有较大的技术发展空间。

5 学术图表知识发现的研究展望

综合学术图表知识发现的技术框架与应用，学术图表知识发现在学术图表对象及文本的识别与获取、学术图表信息表示及标注、学术图表分类和文本分类、学术图表信息抽取等方面取得一定进展，同时在多个应用领域已有相关实践。为使学术图表知识发现能在未来学术知识服务体系中发挥更大作用，本文提出以下几方面的发展策略：

5.1 重视学术图表知识发现，将其融入文献知识发现体系内

长久以来，文本知识发现是文献知识发现的主要实现途径。在全新数据密集型科学发现的科技创新生态下，研究人员愈发重视学术图表，他们迫切需要将学术图表知识发现融入现有文献知识发现。

学术图表知识发现能有效推动文献知识发现服务以适应数据密集型科学发现下的新型知识生态环境。首先对学术图表实施知识发现能扩展现有学术知识检索的对象类型，突破文本检索限制，提供多维学术图表形式的更丰富的文献知识展示，还可通过学术图表发现扩展到学术图表依附的科学数据发现；其次通过学术图表发现实现基于证据的知识精准发现，推动文献知识服务向多模态知识服务进展；最后，基于学术图表信息抽取及学术图表标注，计算机对学术图表理解更深，为深度知识关系挖掘奠定基础。

将学术图表知识发现纳入文献知识发现体系，具体而言：①基于本体学习、本体集成、本体对齐等知识单元语义关联的知识组织方法，构建以学术图表为核心对象的知识单元语义表示模型，如通用描述知识单元、学术图表领域知识单元、面向特定问题解决的知识单元（如自动问答）等，在语义知识组织框架帮助下建设专门的学术图表语义知识库；②利用深度学习、神经网络学习等方法，突破学术图表统一语义表示、学术图表自动分类、学术图表自动语义标注、基于内容的学术图表智能推荐计算、学术图表知识抽取、学术图表自动摘要等关键技术，构建适用于多模态对象的知识发现引擎；③提供创新的针对学术图表特性的不同问题解决的应用组件，例如学术图表语义标注、学术图表自动摘要、学术图表相似检测、学术图表智能问答等，以便研究人员根据自身需求实施数据挖掘和关联。

5.2 完善学术图表语义知识组织体系，构建专门的学术图表语义知识库

语义知识库融合了知识发现技术和知识组织内容，它为新的命名实体识别、语义相似度计算、信息抽取等知识发现技术提供语义数据支撑。在文本知识发现领域，语义知识库已经比较成熟，在领域应用上亦有大量实证。反观学术图表领域，尽管有部分语料库及学术图表数据仓储可供使用，但在学术图表语义知识库上尚处于初步阶段。

构建以学术图表为核心对象的知识单元语义表示模型势在必行。目前学术图表知识组织以传统元数据组织方式为主，以本体和知识学术图表为代表的语义知识模型在学术图表领域正处于新兴发展阶段。从前述学术图表知识发现技术要点来看，现有学术图表知识发现中较少借助知识组织体系，未发挥其在信息检索、信息抽取、实体与关系类型过滤等知识发现过程的作用，这使得学术图表知识发现难以在大规模数据上取得较好效果，也限制学术图表知识发现在领域中的应用。

本体语义知识模型能充当基础知识库中语义类别及关联的框架支撑，同时它在整个语义知识服务的检索到问答过程中发挥语义归一、语义消歧的重要作用。因此需要以学术图表为核心对象，构建适用于不同类型、不同领域、不同问题解决的学术图表知识单元语义表示模型和知识属性体系，采用各类知识单元语义关联的知识组织方法，构建学术图表应用本体、领域本体、知识图谱等。基于语义表示模型应用语义标注技术，建设学术图表基础语料库和知识库。

5.3 以点带面，开发新型学术图表知识发现应用

知识服务是知识发现的价值体现，而知识发现应用是学术图表知识服务快速融入、快速扩展的实现途径。

学术图表检索应用是学术图表知识服务的基础和优先选择。目前国内外相关数字学术提供商如 PMC、ProQuest、CNKI 等，都以学术图表检索为切入点推广学术图表知识服务应用。学术图表检索应结合语义知识组织、检索结果的多重因子排序、智能推荐计算等技术打造学术图表语义智能发现引擎。

学术图表自动摘要不仅是辅助科研人员快速掌握文献内核心内容、筛选所需文献的重要服务，也是将整

chinaXiv:202304.00405v1

体文献知识转换为自然语言表述的模块化知识的重要支撑,进而支撑模块化知识重组、个性化知识串联服务。

问答与推理服务不仅是人机交互中的智能知识服务,还能成为科技创新的重要支撑。学术图表智能问答及推理能够大范围扩展学术知识在日常生活应用,例如基于地区历史日照或降雨统计学术图表,不仅能预测气象,还能提供农作物形态、产量、虫害发生等内容和日照和降雨关联。

近年来,学术图表的不当使用成为学术诚信领域关注的焦点之一。学术图表查重在学术诚信领域大有建树。基于学术图表的细粒度语义标注及图像视觉相似度计算等技术,构建学术图表查重系统,能在一定程度上防止学术图表的不正规重用及数据造假。

6 结语

科技资源深度融合背景下,学术图表知识发现是完善文献知识发现的重要一环,提供除文本知识发现外新的知识发现方式。当前学术图表对象及文本的识别与获取、学术图表信息表示及标注、学术图表分类和文本分类、学术图表信息抽取等学术图表知识发现关键技术日渐成熟,学术图表语义检索、学术图表摘要、学术图表知识问答等新型知识发现服务也正逐步开展。面向未来,我们应当完善学术图表语义知识组织体系、构建专门的学术图表语义知识库、加快开发新型学术图表知识发现应用、推动学术图表知识抽取升级,从而提升科研人员科学发现及知识创造效能,推动数字图书馆知识服务升级。

chinaXiv:202304.00405v1

参考文献:

[1] SIEGEL N, LOURIE N, POWER R, et al. Extracting Scientific figures with distantly supervised neural networks[C]//Proceedings of the 18th ACM-IEEE on joint conference on digital libraries. Texas: ACM,2018: 223-232.

[2] YU H, LEE M. Accessing bioscience images from abstract sentences[J]. Bioinformatics, 2006, 22(14): 547-556.

[3] STELMASZEWSKA H, BLANDFORD A. From physical to digital: a case study of computer scientists' behaviour in physical libraries [J]. International journal on digital libraries, 2004, 4(2): 82-92.

[4] LEE P, WEST J D, HOWE B, et al. Viziometrics: analyzing visual information in the scientific literature[J]. IEEE transactions on big data, 2018, 4(1): 117-129.

[5] PYREDDY P, CROFT W B. TINTIN: A system for retrieval in text tables [C]//Proceedings of the second ACM international conference on digital libraries. Philadelphia: ACM,1997: 193-200.

[6] LIU F, JENSSEN T, NYGAARD V, et al. FigSearch: a figure legend indexing and classification system. [J]. Bioinformatics, 2004, 20(16):2880-2882.

[7] TENOPIR C, SANDUSKY R, CASADO M. The value of CSA deep indexing for researchers (executive summary)[J]. School of information sciences publications and other works, 2006(1): 1-4.

[8] LIU Y, BAI K, MITRA P, et al. TableSeer: automatic table metadata extraction and searching in digital libraries[C] //Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2007:91-100.

[9] XU S H, JAMES M C, MICHAEL K. Yale image finder (YIF) [J]. Bioinformatics, 2008,17(24): 1968-1970.

[10] HONG Y, LIU F, RAMESH B P. Automatic figure ranking and user interfacing for intelligent figure search[J]. Plos one, 2010,5(10): e12983.

[11] NCBI. PMC[EB/OL]. [2020-08-31]. https://www. ncbi. nlm. nih. gov/pmc/.

[12] CNKI. CNKI 图片检索[EB/OL]. [2020-08-31]. http://image. cnki. net/Default. aspx.

[13] SIEGEL N, HORVITZ Z, LEVIN R, et al. FigureSeer: parsing result-figures in research papers [C]//European conference on computer vision. Amsterdam: Springer International Publishing, 2016: 664-680.

[14] National Library of Medicine. Open-i[EB/OL]. [2020-08-31]. https://openi. nlm. nih. gov/.

[15] FAYYAD U M, PIATETSKY-SHAPIRO G, SMYTH P. From data mining to knowledge discovery in databases [J]. Ai magazine, 1996,17(3):37-54.

[16] 唐皓瑾. 一种面向 PDF 文件的表格数据抽取方法的研究与实现[D]. 北京:北京邮电大学, 2015.

[17] 刘颖. 基于 Web 结构的表格信息抽取研究[D]. 合肥:合肥工业大学,2012.

[18] CHAO H, FAN J. Layout and content extraction for PDF documents[C]// Document analysis systems 2004. Florence: Springer, 2004:213-224.

[19] CHOUDHURY S R, GILES C L. An architecture for information extraction from figures in digital libraries[C]//International conference. international world wide web conferences steering committee. Florence: ACM,2015: 667-672.

[20] CHHATKULI A, FONCUBIERTA-RODRÍGUEZ A, MARKONIS D, et al. Separating compound figures in journal articles to allow for subfigure classification[C]//Medical imaging 2013: Advanced pacs-based imaging informatics and therapeutic applications. Florida: SPIE Medical Imaging,2013: 86740J.

[21] LI P, JIANG X, KAMBHAMETTU C, et al. Compound image segmentation of published biomedical figures [J]. Bioinformatics, 2018, 34(7):1192-1199.

[22] Apache Software Foundation. Apache PDFBox[EB/OL]. [2021-05-02]. https://pdfbox. apache. org.

[23] YUSUKE S. PDFMiner[EB/OL]. [2021-05-02]. https://github. com/euske/pdfminer.

[24] Glyph & Cog. Xpdf[EB/OL]. [2021-05-02]. http://www. xpdfreader. com.

ChinaXiv合作期刊

[25] KristianHøgsberg. Poppler[EB/OL]. [2021-05-02]. http://poppler.freedesktop.org/.

[26] LUIS D L, JINGYI Y, CECILIA N, et al. An automatic system for extracting figures and captions in biomedical pdf documents [C]//2011 IEEE international conference on bioinformatics and biomedicine. Atlanta: IEEE, 2011:578-581.

[27] PRACZYK P A, NOGUERAS-ISO J, MELE S. Automatic extraction of figures from scientific publications in high-energy physics [J]. Information technology and libraries, 2013, 32(4):25-52.

[28] CLARK C, DIVVALA S. PDFFigures 2.0: mining figures from research papers [C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. Newark: ACM, 2016:143-152.

[29] LI P Y, JIANG X Y, SHATKAY H, et, al. Figure and caption extraction from biomedical documents. [J]. Bioinformatics, 2019, 35(21):4381-4388.

[30] YILDIZ B, KAISER K, MIKSCH S. Pdf2table: a method to extract table information from pdf files[C]//Proceedings of the 2nd Indian international conference on artificial intelligence. Pune: DBLP, 2008:1-13.

[31] 李海涛, 柳健, 明德烈,等. 一种统计特征点网格分布的表格图像识别方法[J]. 华中科技大学学报(自然科学版), 2002, 30(9):60-63.

[32] 张伯. 基于 PDF 文字流的表格识别技术的研究[D]. 北京:北京工业大学, 2010.

[33] MANUELA, MIKE T, JEREMY B M. Tabula[EB/OL]. [2021-08-31]. https://tabula.technology/.

[34] RASTAN R, PAIK H Y, SHEPHERD J. TEXUS: A unified framework for extracting and understanding tables in PDF documents[J]. Information processing & management, 2019, 55(3): 895-918.

[35] PEREZARRIAGA M O, ESTRADA T, ABADMOTA S. TAO: system for table detection and extraction from pdf documents[C]//Proceedings of the 29th international Florida artificial intelligence research society conference. Florida: AAAI, 2016: 591-596.

[36] SAS J, ZOLNIEREK A. Three-stage method of text region extraction from diagram raster images[J]. Advances in intelligent systems and computing, 2013, 226:527-538.

[37] FALK BÖSCHEN, ANSGAR SCHERP. A Comparison of approaches for automated text extraction from scholarly figures[C]//International conference on multimedia modeling. Reykjavik: Springer, 2017:15-27.

[38] CHIANG Y Y, KNOBLOCK C. A. Recognizing text in raster maps [J]. Geoinformatica, 2015(19):1-27.

[39] XU, S H, MICHAEL K. A new pivoting and iterative text detection algorithm for biomedical images[M]. Elsevier Science, 2010.

[40] DE S, STANLEY R J, CHENG B, et al. Automated text detection and recognition in annotated biomedical publication images[J]. International journal of healthcare information systems and informatics, 2014, 9(2): 34-63.

[41] HE F, WANG D, INNOKENTEVA Y, et al. Extracting molecular entities and their interactions from pathway figures based on deep learning[C]//2019 IEEE international conference on bioinformatics and biomedicine (bibm). San Diego: IEEE, 2020:1191-1193.

[42] NAGY G. Learning the characteristics of critical cells from web tables[C]//International conference on pattern recognition. Tsukuba: IEEE, 2012: 1554-1557.

[43] SETH S C, NAGY G. Segmenting tables via indexing of value cells by table headers[C]//International conference on document analysis and recognition. Washington, DC: IEEE, 2013: 887-891.

[44] HONG Y, AGARWAL S, JOHNSTON M. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension[J]. Journal of biomedical discovery & collaboration, 2009, 4(1):1-10.

[45] CHOUDHURY S R, MITRA P, KIRK A, et, al. Figure metadata extraction from digital documents[C]//International conference on document analysis & recognition. ieee computer society. Washington, DC: IEEE, 2013:135-139.

[46] LOPEZ L D, YU J, ARIGHI C N, et al. An automatic system for extracting figures and captions in biomedical pdf documents[C]//IEEE international conference on bioinformatics & biomedicine. Atlanta: IEEE, 2012:578-581.

[47] BALAJI P R, SETHI R J, HONG Y, et al. Figure-associated text summarization and evaluation [J]. Plos One, 2015, 10(2): e0115671.

[48] YU H. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles[C]//Annual symposium proceedings / amia symposium. amia symposium. Washington, DC: AMia, 2006: 834-838.

[49] BHATIA S, MITRA P. Summarizing figures, tables and algorithms in scientific publications to augment search results[J]. ACM transactions on information systems, 2010, 30(1):1-24.

[50] MANNING C D, RAGHAVAN P, H SCHÜTZE. Introduction to information retrieval [M]. 北京:人民邮电出版社, 2010.

[51] TURTLE H R, CROFT W B. Inference networks for document retrieval[C]//13th international conference on research and development in information retrieval. Brussels: ACM, 1990: 1-24.

[52] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer science, 2013, arXiv:1301.3781.

[53] SHUAI Z, CHENG M M, WARRELL J, et al. Dense semantic image segmentation with objects and attributes[C]//2014 IEEE conference on computer vision and pattern recognition (CVPR). Columbus: IEEE, 2014, 3214-3221.

[54] VEZHNEVETS A, FERRARI V, BUHMANN J. M. Weakly supervised structured output learning for semantic segmentation[C]//2012 IEEE conference on computer vision and pattern recognition. Providence: IEEE, 2012: 845-852.

[55] HUI Z, FRITTS J E, GOLDMAN S A. Image segmentation evaluation: a survey of unsupervised methods[J]. Computer vision & image understanding, 2008, 110(2):260-280.

[56] PEDERSEN K S, LOOG M, DORST P. Salient point and scale de-

chinaXiv:202304.00405v1

tection by minimum likelihood[C]//Proceedings of machine learning research. Bletchley Park: PMLR, 2007: 59 – 72.

[57] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2):91 – 110.

[58] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE computer society conference on computer vision & pattern recognition. San Diego: IEEE, 2005,886 – 893.

[59] NG R T, SEDIGHIAN A. Evaluating multidimensional indexing structures for images transformed by principal component analysis [C]//Proceedings volume 2670, storage and retrieval for still image and video databases iv. San Jose: SPIE, 1996:50 – 61.

[60] PHAM, T T, MAILLOT N E, LIM J H, et al. Latent semantic fusion model for image retrieval and annotation[C]//Proceedings of the sixteenth ACM conference on information and knowledge management. Lisbon: ACM, 2007:439 – 444.

[61] INDYK P. Approximate nearest neighbors: towards removing the curse of dimensionality[C]//Proceedings of the 30th acm symposium on theory of computing (stoc'98). Dallas Texas: ACM, 1998: 604 – 613.

[62] 杨战波. 基于深度学习和词嵌入的视觉语义嵌入研究[D]重庆:西南大学,2019.

[63] WANG H, ZHANG Y, JI Z, et al. Consensus-aware visual-semantic embedding for image-text matching[C]// 2020 european conference on computer vision. Glasgow: Qrxiv, 2020: 18 – 34.

[64] WEN K, GU X, CHENG Q. Learning dual semantic relations with graph attention for image-text matching[J]. IEEE transactions on circuits and systems for video technology, 2020 (99):1 – 1.

[65] 陈涛,单蓉蓉,李惠. 数字人文中图像资源的语义化标注研究[J]. 农业图书情报学报,2020,32(9): 6 – 14.

[66] BHAGAT P K, CHOUDHARY P. Image annotation: then and now [J]. Image and vision computing, 2018(80): 1 – 23.

[67] ADNAN M M, RAHIM M, REHMAN A, et al. Automatic image annotation based on deep learning models: a systematic review and future challenges[J]. IEEE access, 2021(9): 50253 – 50264.

[68] MIAO R, TOTH R, ZHOU Y, et al. Quick annotator: an open-source digital pathology based rapid image annotation tool[J] The journal of pathology,20217(6): 542 – 547.

[69] DONG Q, LUO G, HAYNOR D, et al. DicomAnnotator: a configurable open-source software program for efficient dicom image annotation[J]. Journal of digital imaging,2020,33(6): 1514 – 1526.

[70] 孙坦, 丁培, 黄永文, 等. 文本挖掘技术在农业知识服务中的应用述评[J]. 农业图书情报学报, 2021, 33(1): 4 – 16.

[71] POCO J, HEER J. Reverse - engineering visualizations: recovering visual encodings from chart images[J]. Computer graphics forum, 2017, 36(3):353 – 363.

[72] KIM S, LIU Y. Functional – based table category identification in digital library[C]//2011 international conference on document analysis and recognition, ieee,2011: 1364 – 1368.

[73] SAVVA M, KONG N, CHHAJTA A, et al. ReVision: automated classification, analysis and redesign of chart images[C]//User interface software and technology. New York: ACM, 2011: 393 – 402.

[74] NKWENTSHA X, HOUNKANRIN A, NICOLLS F. Automatic classification of medical X-ray images with convolutional neural networks[C]//2020 international saupec/robmech/prasa conference. Cape Town: Springer, 2020:1 – 4.

[75] HUANG W, ZONG S, TAN C L, et al. Chart image classification using multiple-instance learning[C]//Workshop on applications of computer vision. Texas: ACM, 2007: 27 – 27.

[76] PELKA O, FRIEDRICH C M. FHDO biomedical computer science group at medical classification task of ImageCLEF 2015 [C]// Working notes of CLEF 2015 conference. Toulouse: CEUR-WS, 2015.

[77] LI P,SORENSEN S,KOLAGUNDA A, et al. UDEL CIS working notes in ImageCLEF 2016 [C]//Working notes of CLEF 2016 conference. Portugal: CEUR-WS, 2016:334 – 346.

[78] CHHATKULI A, FONCUBIERTA-RODRIGUEZ A, MARKONIS D, et al. Separating compound figures in journal articles to allow for subfigure classification [C]//Proceedings of spie medical imaging, advanced pacs-based imaging informatics and therapeutic applications. Orlando: SPIE, 2013:86740.

[79] YUAN X, ANG D. A novel figure panel classification and extraction method for document image understanding [J]. International journal of data mining and bioinformatics, 2014, 9(1):22 – 36.

[80] Li P, Jiang X, Kambhamettu C, et al. Segmenting compound biomedical figures into their constituent panels[C]//International conference of the cross-language evaluation forum for europeanm languages. Dublin: Springer, 2017:199 – 210.

[81] TASCHWER M, MARQUES O. Compound figure separation combining edge and band separator detection[C]//International conference on multimedia modeling. Miami: Springer, 2016:162 – 173.

[82] SANTOSH K C, AAFAQUE A, ANTANI S, et al. Line segment-based stitched multipanel figure separation for effective biomedical CBIR [J]. International journal of pattern recognition and artificial intelligence, 2017, 31(6):1757003.

[83] 于玉海. 面向医学文献的图像模式识别关键技术研究[D]. 大连:大连理工大学. 2018.

[84] CRESTAN E, PANTEL P. Web-scale table census and classification[C]//Proceedings of the fourth acm international conference on web search and data mining. Hong Kong: ACM,2011: 545 – 554.

[85] MURPHY R F, VELLISTE M, YAO J, et al. Searching online journals for fluorescence microscope images depicting protein subcellular location patterns[C]//IEEE international symposium on bioinformatics & bioengineering. Bethesda: IEEE, 2001:119 – 128.

[86] GERTZ M, SATTLER K U, GORIN F, et al. Annotating scientific images: a concept-based approach[C]//Proceedings 14th international conference on scientific and statistical database management. Los Alamitos: IEEE, 2002:59 – 68.

[87] EMAGE. Data Annotation Methods [EB/OL]. [2020 – 11 – 02]. http://www. emouseatlas. org/emage/about/data _ annotation _

chinaXiv:202304.00405v1

ChinaXiv合作期刊

methods. html#auto_eurexpress.

[88] TOO E C, YUJIAN L, NJUKI S, et al. A comparative study of fine-tuning deep learning models for plant disease identification [J]. Computers and electronics in agriculture, 2018,161(1):272 –279.

[89] BARBEDO J A. Plant disease identification from individual lesions and spots using deep learning[J]. Biosystems engineering, 2019, 180(1):96 –107.

[90] KUHN T, NAGY M, LUONG T B, et al. Mining images in biomedical publications: Detection and analysis of gel diagrams[J]. J biomed semantics, 2014, 5(1):1 –9.

[91] ZHANG Z. Towards efficient and effective semantic table interpretation[C]//International semantic Web conference. New York: Springer-verlag, 2014:487 –502.

[92] CAO H, BOWERS S, SCHILDHAUER M P. Approaches for semantically annotating and discovering scientific observational data [C]//Database and expert systems applications. Berlin: Springer, 2011: 526 –541.

[93] MARTIN M, NUFFELEN B, ABRUZZINI S, et al. The digital agenda scoreboard: a statistical anatomy of Europe's way into the information age[EB/OL]. [2021 –05 –02]. http://www.semantic-web-journal.net/sites/default/files/swj283.pdf.

[94] KEMBHAVI A, SALVATO M, KOLVE E, et al. A diagram is worth a dozen images[C]//Computer vision - eccv 2016. Amsterdam: Springer, 2016: 235 –251.

[95] LEE P, YANG T. S, WEST J, et al. Phyloparser: a hybrid algorithm for extractingphylogenies from dendrograms[C]//14th iapr international conference on document analysis and recognition (icdar). Kyoto: IEEE, 2017:1087 –1094.

[96] 何英. PubMed Central 文献中的柱形图信息抽取研究与应用[D]. 武汉:武汉理工大学,2018.

[97] AGARWAL S, YU H. FigSum: automatically generating structured text summaries for figures in biomedical literature. [C]//American medical informatics association annual symposium. San Francisco: PMC, 2009: 6 –10.

[98] SAINI N, SAHA S, POTNURUV, et al. Figure summarization: a multiobjective optimization-based approach [J]. Intelligent systems, 2019,34(6):43 –52.

[99] SAINI N, SAHA S, BHATTACHARYYA P, et al. Textual entailment——based figure summarization for biomedical articles[J]. ACM transactions on multimedia computing communications and applications, 2020, 16(1s):1 –24.

[100] CHEN J, ZHUGE H. Extractive summarization of documents with images based on multi-modal RNN[J]. Future generation computer systems, 2019,99(1): 186 –196.

[101] 吴晨飞. 基于关系建模的视觉问答研究[D]. 北京:北京邮电大学, 2020.

[102] KAFLE K, PRICE B, COHEN S, et al. DVQA: understanding data visualizations via question answering[C]//2018 IEEE/cvf conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018:5648 –5656.

[103] KAHOU S E, MICHALSKI V, ATKINSON A, et al. FigureQA: an annotated figure dataset for visual reasoning[J]. Computer science, 2018, arXiv:1710.07300.

[104] CHAUDHRY R, SHEKHAR S, GUPTA U, et al. LEAF-QA: locate, encode & attend for figure question answering[C]// 2020 IEEE winter conference on applications of computer vision (wacv). Snowmass Village: IEEE, 2020: 3512 –3521.

chinaXiv:202304.00405v1

The Technical Framework and Research Progress of Knowledge Discovery in Academic Figures and Tables

Ding Pei
Shenzhen University Library, Shenzhen 518060

Abstract: [**Purpose/significance**] Under the background of deep integration of scientific resources, knowledge discovery of academic figures and tables provides a new way of knowledge discovery besides text knowledge discovery. Knowledge discovery of academic figures and tables is an important segment in document knowledge discovery perfection, it improves the efficiency of scientific discovery and knowledge creation of researchers and promotes the upgrade of knowledge service of digital library. [**Method/process**] This paper sort out the evolution of knowledge discovery of academic figures and tables, demonstrated its technical framework in detail and proved that the knowledge discovery technology of academic figures and tables had been gradually mature. Combined with knowledge discovery application service with academic charts, this paper found that knowledge discovery of academic figures and tables could support scientific and technological innovation activities in many ways. [**Result/conclusion**] Looking into the future, we need to: attach importance to the knowledge discovery of academic figures and tables and integrate it into the literature knowledge discovery system; perfect the semantic knowledge organization system of academic figures and tables and build a special semantic knowledge base of academic figures and tables; develop new knowledge discovery applications for academic figures and tables.

Keywords: academic figures and tables knowledge discovery knowledge organization information extraction